TextCritical.net - Feature #2783

Improved lemmatization

06/13/2020 05:39 PM - Luke Murphey

Status:	Closed	**Start date:**	06/13/2020
Priority:	Normal	**Due date:**	
Assignee:	Luke Murphey	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			

Description

See https://github.com/PhVerkerk/Eulexis_off_line/tree/master/Eulexis_data

- https://github.com/PhVerkerk/Eulexis_off_line/blob/master/Eulexis_data/analyses_gr.txt
- https://github.com/PhVerkerk/Eulexis_off_line/blob/master/Eulexis_data/LSJ.csv

History

#1 - 06/13/2020 05:52 PM - Luke Murphey

Obs:

- https://raw.githubusercontent.com/PhVerkerk/Eulexis_off_line/master/Eulexis_data/analyses_gr.txt has parses but no short definitions
- https://raw.githubusercontent.com/PhVerkerk/Eulexis_off_line/master/Eulexis_data/LSJ.csv has links to LSJ
- https://github.com/PhVerkerk/Eulexis_off_line/blob/master/Eulexis_data/trad_gr_en_fr_de.csv has short definitions in multiple languages
  - This project is related to https://github.com/pjheslin/diogenes
- Perseus has an API too: http://sites.tufts.edu/perseusupdates/2012/11/01/morphology-service-beta/
  - http://services.perseids.org/bsp/morphologyservice/analysis/word?lang=grc&engine=morpheusgrc&word=%E1%BC%B1%CF%83%CF%84%CE%BF%CF%81%CE%AF%CE%B7%CF%82

Qs:

- Is the trad_gr_en_fr_de.csv better that the one I have?
- How would I want to use these online tools?
  - Have a UI option to parse it
  - Have the UI parse it when it cannot find a parse
  - Have the UI parse it when it cannot find a LSJ definition
- Where is the Diogenes list? How does it compare?
  - https://github.com/PhVerkerk/Eulexis_off_line/tree/master/Eulexis_data
  - https://raw.githubusercontent.com/PhVerkerk/Eulexis_off_line/master/Eulexis_data/trad_gr_en_fr_de.csv
- What does the Diogenes data look like?
  - https://github.com/pjheslin/diogenes/blob/master/mk.prebuilt-data
  - https://github.com/pjheslin/diogenes-prebuilt-data/raw/master/prebuilt-data.tar.xz
    - https://github.com/pjheslin/diogenes-prebuilt-data
- What are some strange parses I ought to consider as test cases?
  - https://textcritical.net/work/new-testament/Luke
    - καὶ
    - γενόμενοι
    - ἔδοξε
    - κἀμοὶ
    - καθεξῆς
    - λόγων
    - τὴν
    - ΕΓΕΝΕΤΟ
    - στεῖρα

#2 - 06/15/2020 05:01 AM - Luke Murphey

- Target version deleted (4.2.0)

#3 - 12/20/2020 05:14 AM - Luke Murphey

- Target version set to 4.6.7

https://textcritical.net/work/antiquitates-judaicae

- Τοῖς
  - lentil
- τὰς
  - lentil
- ἱστορίας
- συγγράφειν
  - ἀντι-συγγραφω (pres inf act): write
- βουλομένοις
  - ἀνα-βουλομαι (pres part mid-pass masc/nuet dat pl): will
- οὐ
  - οὐ (adverbial indeclinable): in truth
- μίαν
  - εἷς (fem acc sg): sem
- οὐδὲ
  - οὐδοσ1 (masc voc sg): threshold
- τὴν
- αὐτὴν
  - ἀυτη (fem acc sg): cry
- ὁρῶ
  - ἀμφι-ὁραω (pres mid imperat 2nd sg): Inscr. destombeaux des rois
- τῆς
- σπουδῆς
  - ἀνα-σπουδαζω (fut act ind 2nd sg): to be busy
- γινομένην
  - ἀνα-γιγνομαι (pres part mid-pass fem acc sg): come into a new state of being
- αἰτίαν,
- ἀλλὰ
  - ἄλλος (nuet nom/voc/acc pl): y
- πολλὰς
- καὶ
  - ἀντι-καιω (pres act imperat 2nd sg): kindle
- πλεῖστον
- ἀλλήλων
- διαφερούσας.

#4 - 12/20/2020 05:37 AM - Luke Murphey

I lemmatized "Τοῖς τὰς ἱστορίας συγγράφειν βουλομένοις οὐ μίαν οὐδὲ τὴν αὐτὴν ὁρῶ τῆς σπουδῆς γινομένην αἰτίαν, ἀλλὰ πολλὰς καὶ πλεῖστον ἀλλήλων διαφερούσας" at https://outils.biblissima.fr/en/eulexis-web/

This doesn't look much better than what I have.

#5 - 12/20/2020 05:41 AM - Luke Murphey

There is a lemmatizer here: https://docs.cltk.org/en/latest/greek.html

This is based on greek_models_cltk (https://github.com/nodage/greek_models_cltk)

Import it here: https://docs.cltk.org/en/latest/importing_corpora.html

```
from cltk.tag.pos import POSTag
tagger = POSTag('greek')
tagger.tag_ngram_123_backoff('θεοὺς μὲν αἰτῶ τῶνδ᾽ ἀπαλλαγὴν πόνων φρουρᾶς ἐτείας μῆκος')
```

It also parses TEI: https://docs.cltk.org/en/latest/greek.html#tei-xml

#6 - 12/20/2020 05:45 AM - Luke Murphey

https://github.com/GreekPerspective/glem

#7 - 12/20/2020 05:53 AM - Luke Murphey

See also https://github.com/stenskjaer/lemmatizer

#8 - 12/20/2020 06:40 AM - Luke Murphey

```
from cltk.corpus.utils.importer import CorpusImporter
corpus_importer = CorpusImporter('greek')
corpus_importer.import_corpus('greek_models_cltk')
```

#9 - 12/29/2020 02:00 AM - Luke Murphey

Another source:

- http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0007&redirect=true
- http://dcc.dickinson.edu/greek-core-list
- https://jktauber.com/2015/11/16/actual-core-vocab-lists-greek-new-testament/

#10 - 01/01/2021 11:24 PM - Luke Murphey

- Target version deleted (4.6.7)

#11 - 03/25/2022 04:58 AM - Luke Murphey

- Status changed from New to Closed

- % Done changed from 0 to 100